图书情报工作
第63卷 第9期 2019年5月


# 专利被引频次的时间影响研究*

■ 罗文馨[1,2] 赵亚娟[1,2]
[1] 中国科学院文献情报中心 北京 100190
[2] 中国科学院大学经济管理学院图书情报与档案管理系 北京 100190

**摘要：**[**目的/意义**]研究时间因素对专利被引频次的影响,可以减少时间因素对技术评价活动的制约,提高评价的准确性和可信度。[**方法/过程**]采集1975－2017年的美国专利数据,开展基于固定效应法的专利被引频次的修正研究。将专利按照不同公开年份和不同技术领域分组,选定组内均值和6个TOP分位数为被引频次基准,统计当前时间点的被引频次基准线及基准线的历史时序变化情况。建立神经网络模型,拟合基准线的时序变化规律,并预测未来统计时间点的基准线。[**结果/结论**]专利公开年份和统计年份的时间差异,使得专利被引频次无法直接进行比较。本文建立了基于不同技术领域、不同公开年份和不同统计年份的专利被引频次基准线,为专利评估提供参考。

**关键词：**专利被引频次 时间 固定效应 时序变化 基准
**分类号：**G250.2
**DOI**:10.13266/j.issn.0252－3116.2019.09.006



## 1 引言

专利引文分析是专利分析的重要内容,对于识别重要专利、探索技术发展路线以及评估专利价值等具有重要意义。专利引文分析依赖一系列专利引文测度指标,例如,引文数量和被引频次等。在实际应用中,因专利引文产生方式的多样性和专利引文所揭示含义的丰富性,专利引文测度指标的合理使用有待深入研究[1]。本文聚焦于专利引文测度指标的时间影响研究。

时间因素对引文测度指标的影响,一直受到国内外学者们的关注[2－3],多数研究聚焦于论文引文测度指标的时间影响研究。在论文引文研究的基础上,建立了学术界广泛认可的ESI(Essential Science Indicators)体系,提供了基于不同领域、不同时间范围的引文指标基准值,用于消除时间和领域因素对论文引文评价指标的影响。但是在专利领域,缺少类似实践工作,尚未形成基于不同领域、不同时间范围的基准,使得专利引文测度缺少标准,无据可依。

专利引文测度指标包括两个基本指标和一系列衍生指标。两个基本指标为引文数量和被引频次。其中,引文数量可以细分为专利引文数量和非专利引文数量,衍生出技术原创性、技术普遍性等指标。被引频次则衍生出引证指数、科学关联性、技术实力等指标。专利引文测度指标受时间影响,主要是由于两个基本指标受时间影响。这两个基本指标中,被引频次比引文数量更具现实意义。本文关注专利引文测度指标中的被引频次。

B. H. Hall在2001年通过研究1975－1999年的美国专利引用数据,最早提出专利被引频次的时间影响问题[4]:①任何专利的被引频次都只是截至统计时间的被引情况;②专利引用受专利审查制度的影响,不同时期专利审查制度有差异,导致专利被引机会也有所不同。后续的一些实证研究发现,除美国外,其他国家的专利被引频次也有类似表现。较早期的研究一直停留在时间影响的现象描述上,直到2014年,万小丽[5]受B. H. Hall的研究启发,详细剖析了专利被引频次受时间影响的原因。她举例说明了"时间截面"问题和"引证膨胀"问题:"时间截面"问题指由于缺少未来被引频次,专利被引频次是不完整统计;"引证膨胀"问

* 本文系国家重点研发计划"知识产权信息共享与运营服务应用示范"(项目编号:SQ2017YFB140324)研究成果之一。
作者简介:罗文馨(ORCID:0000－0002－3866－7391),硕士研究生;赵亚娟(ORCID:0000－0003－3501－8131),研究员,博士生导师,通讯作者,E-mail:zhaoyj@mail.las.ac.cn。
收稿日期:2018－10－25 修回日期:2019－01－22 本文起止页码:47－60 本文责任编辑:易飞

题指专利引用其他专利的平均数量逐年增长，单件专利的被引机会越来越大。不仅评价单件专利时存在上述问题，A. Breitzman 在2015年发现，评价一组专利的引用情况时也有类似问题[6]。近年来，相关研究依然保持热度。2017年，A. B. Jaffe 在回顾专利引用的相关社会学研究时，再次强调了专利被引频次会随时间累积不断增加，表现出强烈的队列效应（cohort effects）[1]。队列效应是社会学研究中的重要概念，这里具体是指：不同时期的专利的被引频次受时间影响不同。

梳理相关研究成果后，发现时间对专利被引频次的影响主要在于三个方面：①专利公开时间越接近统计时间点，被引频次越低。不同年份公开专利的被引频次，无法直接进行比较；②计算机技术的快速进步，提高了专利审查员的检索能力，导致专利被引机会逐年增大。因此，无法直接比较在相同时间间隔下，不同年份公开专利的被引频次；③只能计算截至统计时间点的专利被引频次，未来被引频次难以预知，导致评价年轻专利时误差较大。这三个问题均是评价专利的公开年份不同或统计年份不同引起的。

基于以上现状，本文旨在研究时间因素对专利被引频次的影响，以解决专利不同公开年份和不同统计年份的时间差异引起的测度问题。

## 2 数据与方法

### 2.1 数据采集及处理

本文的研究数据从 DI 数据库（Derwent Innovation）中采集。美国专利引用信息更完备，故本文选择美国为研究对象。美国的专利引用数据从1975年开始计算机化，故数据范围限定为美国1975－2017年公开的所有专利记录及其施引专利记录。

研究数据集分为 I 基础专利集合和 II 施引专利集合。I 基础专利集合为主要数据集，即美国1975－2017年公开的所有专利记录。II 施引专利集合为扩展数据集，即 I 基础专利集合中所有专利的施引专利集合。采集的专利著录字段见表1。

数据采集及处理过程如图1所示。采集 I 基础专利数据时，限定检索库为美国授权专利库和美国专利申请库，检索式为“DP >＝(19750101) AND DP <＝(20171231)”，检索时间为2018年1月20日，共得到11 742 361条记录。采集 II 施引专利数据时，先从 I 基础专利集合提取施引专利的公开号，去重后得到公开号列表，用于公开号检索，最终得到19 953 069条记录。

表1 数据集著录字段及含义

| 专利集合 | 字段名 |
|---|---|
| I 基础专利集合：美国1975－2017年公开的所有专利记录 | 1公开号；2标题；3申请号；4申请日期；5公开日期；6公开年；7 IPC 部；8 IPC 现版完整；9 DWPI 分类；10 DWPI 手工代码；11 施引专利；12 施引专利计数；13 施引专利详细信息 |
| II 施引专利集合：基础专利集的施引专利集合 | 1公开号；2公开年 |

数据处理阶段包括数据库构建、缺失值处理、异常记录筛查和数据合并更新等。其中，数据合并更新步骤是：将 I 基础专利集合的专利记录，按照相同申请号字段合并、去重，公开年和公开日期字段保留重复记录中的最早公开年和公开日期，施引专利字段取重复记录中施引专利的并集。数据合并更新的原因是：一项美国专利，从提交申请到授权阶段可能生成多个公开文件，在 DI 数据库中出现多条记录，然而这些记录的被引情况并不是完全一致。

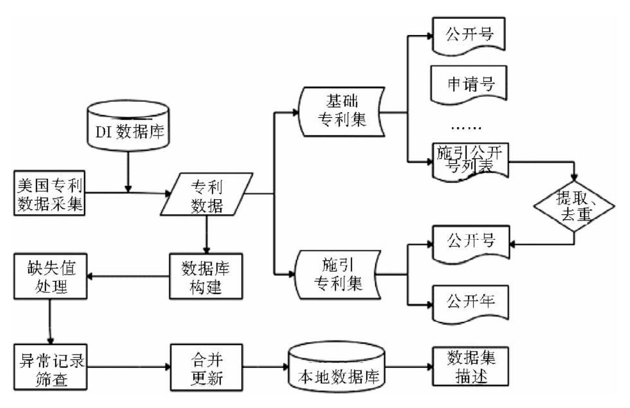



图1 数据采集及处理过程

### 2.2 研究思路及方法

本文研究思路如下：首先，针对专利被引频次受时间影响、缺少评价标准的现状，梳理相关研究，发现时间因素引起专利被引频次无法直接比较的关键问题在于专利不同公开年份和不同统计年份的时间差异。其次，为了消除该时间影响，总结专利被引频次的时间影响的修正方法，筛选出合适的修正方法：百分位数、相对影响指标及固定效应法。最后，在采集处理专利数据的基础上，采用选定的修正方法，开展基于当前时间点及基于时序变化的时间影响修正研究。见图2。

## 3 专利被引频次的时间影响修正

时间因素对引文测度影响的修正方法，可以细分为相对影响指标[7-8]、加入时间因子[9-10]、引证窗口选择[11-12]、百分位数[13-14]、模拟分布曲线[9]和引入非引用指标[15-16]等。在论文领域中，相对影响指标和百分

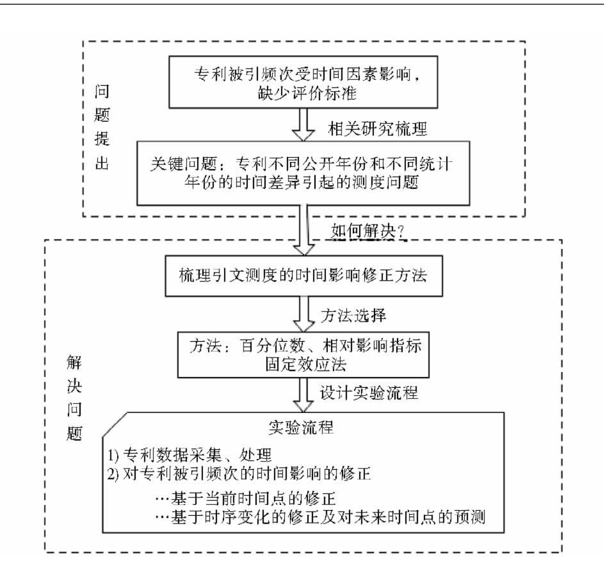



图 2　研究思路



位数方法常被用于多个学科领域和不同指标的修正，适用面很广，其他方法大多应用于某一特定学科领域或特定指标，普适性有待验证。在专利领域，B. H. Hall 针对美国专利商标局受理的 1969－1999 年专利文献，计算了截至 1999 年底的专利被引频次的均值，提出用固定效应方法来进行专利被引频次指标的标准化[6]。他选取同类专利的被引频次均值为基准，提出了专利的相对被引频次，即用一项专利的绝对被引频次除以该专利所在技术领域同年授权专利的平均被引频次。固定效应(fixed effect)最早是实验设计的基本概念，是指在实验中选取某观察因素的水平效应作为固定参数。在统计学中，固定效应是模型参数固定的一种统计模型。例如，将数据按照几个因素分组，选取分组后的组内均值为每个小组的固定效应。

综合论文和专利领域的修正方法，本文采用固定效应方法，选择专利被引频次的均值和 TOP 分位数作为基准，对专利被引频次的时间影响进行修正。假设专利被引频次随公开年份和统计年份的推移引起的变化都是系统性变化，在比较不同专利的被引频次时需要消除这种变化。将专利按照不同公开年份和不同统计年份分组，选取组内均值和 TOP 分位数作为每个小组的固定效应。因为不同技术领域的专利被引情况差别较大，所以分组时还加入了技术领域类别。

修正研究分为基于当前时间点的修正和基于历史时序变化的修正。不仅为当前时间点的专利评估提供基准线，还通过拟合基准线的历史时序变化规律，对未来时间点的基准线进行预测。

## 3.1　基于当前时间点的基准线

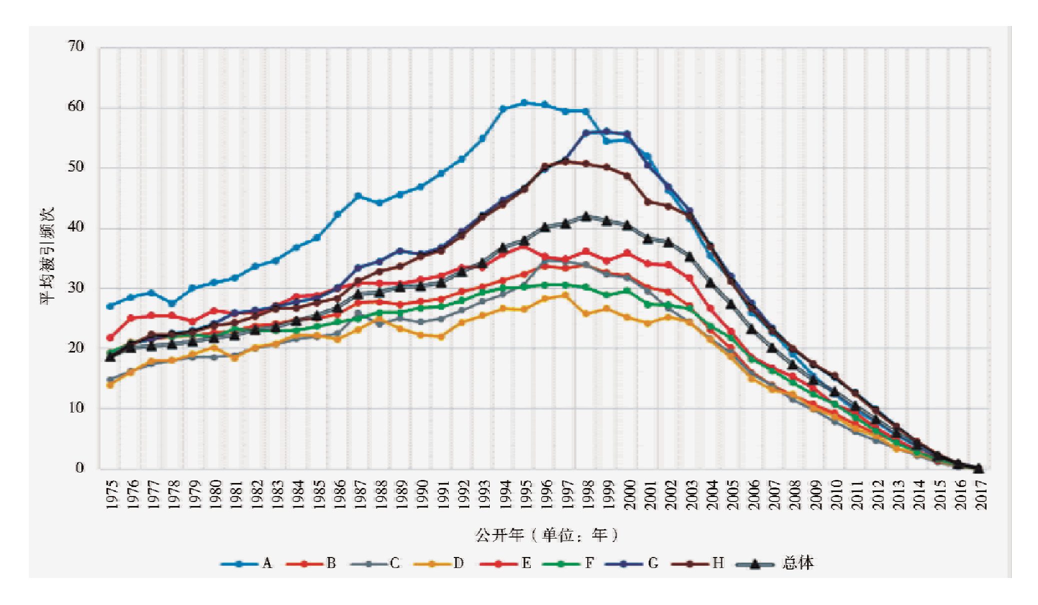
统计时间点为 2017 年底时，专利被引频次按照不同技术领域(IPC 部)和不同公开年分组后，计算的基准线见表 2 和表 3。其中表 2 表示组内均值，表 3 表示组内 TOP 分位数水平。TOP 分位数表示各个 TOP 百分位水平的专利被引频次的值，共选取了 6 个 TOP 分位数，分别为 TOP0.01%(排名前 0.01%)、TOP0.10%(排名前 0.10%)、TOP1.00%(排名前 1.00%)、TOP10.00%(排名前 10.00%)、TOP20.00%(排名前 20.00%)和 TOP 50.00%(排名前 50.00%)。图 3 是表 2 的图形化描述。截至 2017 年底，全领域和不同技术领域专利被引频次的均值，都随着公开年的增长呈现出先增加后下降的趋势。近十几年间公开专利被引频次均值的下降是因为受到统计时间截断的影响，新公开的专利还没有被在后公开的专利充分引用。早些年公开的专利可以近似认为不受统计时间的截断影响，容易发现，不同年份公开的专利的被引情况差别较大。对比不同技术领域的专利被引频次，发现不同技术领域峰值所对应的公开年份有差异，如 A 部(人类生活必需)在公开年为 1996 年时专利被引频次均值达到峰值，而 G 类(物理)则是在公开年为 2000 年时达到峰值。不同技术领域的专利被引频次均值的大小也有差异，如 A 部(人类生活必需)专利的被引频次均值明显高于其他各部，G 部(物理)和 H 部(电学)专利的被引频次均值大小接近，D 部(纺织;造纸)各年公开专利的被引频次均值最低。

评价单件专利时，根据公开年和 IPC 部找到该专利对应的分组，用该专利的被引频次除以表 2 中的组内均值，或者将其与表 3 中 TOP 分位数水平进行比对。可以分别用于修正公开年或技术领域的影响。

评价专利集合时，可以计算该专利集合的被引频次加权值。把专利集合中的专利按公开年和 IPC 部进行分组，找到表 3 中对应小组的 TOP 分位数。按照 TOP 分位数将专利被引频次划分到 7 个百分位区段：高于或等于 TOP0.01%、TOP0.01% －0.10%、TOP0.10% －1.00%、TOP1.00% －10.00%、TOP10.00% －20.00%、TOP20.00% －50.00% 和低于 TOP50.00%，并计算每个百分位区段中专利数占总集合专利数的比例。不同百分位区段赋予不同权重，用百分位区段对应权重乘以该百分位区段的专利数目占比，求和后得到被引频次的加权值。

## 3.2　基于时序变化的基准线及预测

3.2.1　基于历史时序的基准线　专利被引频次随着统计年份的推移而增加。在不同统计年份，被引频次的均值和 TOP 分位数水平不同。前人研究中，大多是




注：图中字母代表 IPC 部，含义分别为：A－人类生活必需；B－作业、运输；C－化学、冶金；D－纺织、造纸；E－固定建筑物、F－机械工程、照明、加热、武器、爆破；G－物理；H－电学。总体代表全领域。下同

图 3　不同公开年份的美国专利被引频次均值变化情况

表 2　不同技术领域下不同公开年份专利的被引频次均值示例

| 公开年 | A 部 | B 部 | C 部 | D 部 | E 部 | F 部 | G 部 | H 部 | 所有类 |
|---|---|---|---|---|---|---|---|---|---|
| 1975 | 27.04 | 19.26 | 14.85 | 14.00 | 21.81 | 19.36 | 18.59 | 18.43 | 18.70 |
| 1976 | 28.50 | 21.03 | 16.30 | 16.09 | 25.04 | 20.90 | 20.78 | 20.59 | 20.27 |
| 1977 | 29.29 | 21.58 | 17.48 | 17.96 | 25.50 | 22.05 | 21.62 | 22.36 | 21.23 |
| … | … | … | … | … | … | … | … | … | … |
| 2015 | 1.94 | 1.61 | 1.09 | 1.43 | 1.52 | 1.54 | 2.21 | 2.35 | 1.96 |
| 2016 | 0.76 | 0.67 | 0.40 | 0.51 | 0.59 | 0.65 | 0.87 | 0.987 | 0.79 |
| 2017 | 0.10 | 0.11 | 0.057 | 0.07 | 0.09 | 0.11 | 0.14 | 0.176 | 0.13 |
| 所有年 | 26.85 | 18.18 | 16.92 | 18.10 | 20.22 | 17.14 | 23.53 | 22.26 | 21.01 |

表 3　不同技术领域下不同公开年份专利的被引频次 TOP 分位数示例

| 公开年 | IPC 部 | TOP0.01% | TOP0.10% | TOP1.00% | TOP10.00% | TOP20% | TOP50% |
|---|---|---|---|---|---|---|---|
| 1975 | A | 1 982.16 | 520.52 | 188 | 58 | 36 | 15 |
| 1975 | B | 521.88 | 265.06 | 106 | 42 | 28 | 13 |
| 1975 | C | 548.85 | 269.49 | 103 | 34 | 21 | 8 |
| 1975 | D | 434.08 | 293.3 | 92.15 | 30.50 | 20 | 8 |
| 1975 | E | 402.17 | 276.97 | 120.44 | 45 | 33 | 15 |
| 1975 | F | 532.71 | 226.68 | 103 | 42 | 29 | 13 |
| 1975 | G | 789.15 | 324.85 | 119 | 41 | 26 | 11 |
| 1975 | H | 564.47 | 263 | 125 | 40 | 26 | 11 |
| 1975 | 全类别 | 1 068.81 | 321.51 | 119 | 41 | 27 | 11 |
| 1976 | A | 924.77 | 510.91 | 222.56 | 62 | 39 | 15 |
| 1976 | B | 485.96 | 254 | 117 | 45 | 31 | 14 |
| … | … | … | … | … | … | … | … |
| 2016 | H | 84 | 41.47 | 10 | 3 | 2 | 0 |
| 2016 | 全类别 | 90.65 | 41 | 10 | 2 | 1 | 0 |
| 2017 | A | 19 | 7 | 2 | 0 | 0 | 0 |
| 2017 | B | 11 | 5 | 2 | 0 | 0 | 0 |
| 2017 | C | 12 | 5 | 2 | 0 | 0 | 0 |
| 2017 | D | 4 | 3.83 | 2 | 0 | 0 | 0 |
| 2017 | E | 6 | 5 | 2 | 0 | 0 | 0 |
| 2017 | F | 12 | 8 | 2 | 0 | 0 | 0 |
| 2017 | G | 15 | 6 | 2 | 0 | 0 | 0 |
| 2017 | H | 30 | 9 | 3 | 1 | 0 | 0 |
| 2017 | 全类别 | 25 | 9 | 3 | 0 | 0 | 0 |

基于当时的统计年份。本文基于不同的统计年份，计算专利被引频次的均值和TOP分位数，探讨专利被引频次基准线的逐年变化情况和增长规律。在此基础上，根据历史数据对基准线进行拟合，预测未来统计年份的基准线，为未来的专利评估提供参考。

图4显示了不同公开年份专利在不同统计年份的被引频次均值分布情况。图5显示了八大技术领域下不同公开年份专利在不同统计年份的被引频次均值分布情况。图4和图5中，X轴表示专利公开年份，Y轴表示专利统计年份，Z轴表示专利被引频次均值。从全领域的被引频次均值变化情况来看，固定公开年不变时，被引频次均值随着统计年的增长呈现上升趋势；固定统计年不变时，被引频次均值随公开年的增长呈先增后减的趋势。1985年之前公开的专利，被引频次均值随统计年增加的增长速率比较平缓；1985－1990年间公开的专利，被引频次均值随统计年增加的增长速率明显加快；1990－2000年间公开的专利，被引频次均值随统计年增加的增长速率更快。在不考虑统计时间截断的条件下，专利公开年份越近，被引频次均值随统计时间增加的增长速率越快。这一现象可能是因为计算机技术的快速进步，使得专利审查员的检索能力提高，从而专利被引机会逐年增大。对比不同领域的被引频次均值变化情况，A部（人类生活必需）、C部（化学；冶金）、G部（物理）、H部（电学）的专利被引频次随公开年份和统计年份变化的形态相似，与全领域的变化规律类似。B部（作业；运输）、F部（机械工程；照明；加热；武器；爆破）的专利，被引频次均值随统计年增加的增长速率变化较为平缓。E部（固定建筑物）的专利在固定专利公开年时，被引频次随统计时间递增的增长速率先慢后快，在统计年为2005年时有明显转折。D部（纺织；造纸）的专利在固定专利统计年时，被引频次均值随公开年递增呈上下波动的趋势，而其他各部的专利被引频次则是随公开年递增呈先增后降的趋势。

除被引频次均值外，被引频次TOP分位数分布示例见图6和图7。图6为不同公开年份专利在不同统计年份的被引频次TOP1.00%分位数分布情况。不同IPC技术领域下，不同公开年份专利在不同统计年份的被引频次TOP分位数分布差异较大。作为示例，图7给出了IPC领域中的A部和B部的TOP 1.00%分位数分布情况。图6和图7中，X轴表示专利公开年份，Y轴表示专利统计年份，Z轴表示专利的TOP1.00%分位数。

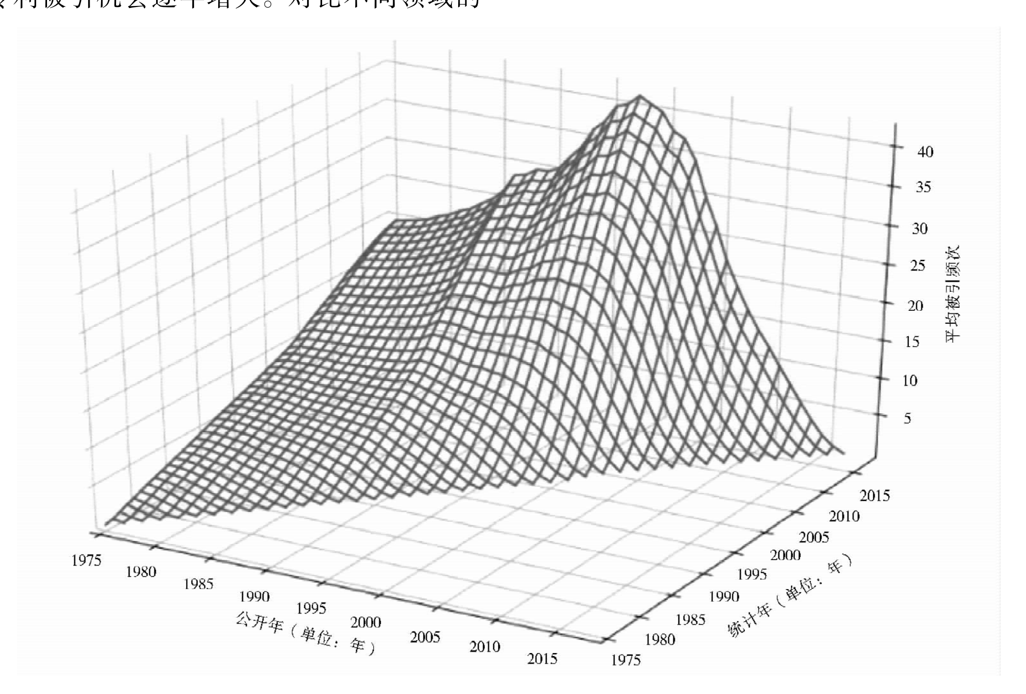



图4　不同公开年份专利截至不同统计时间的被引频次均值变化

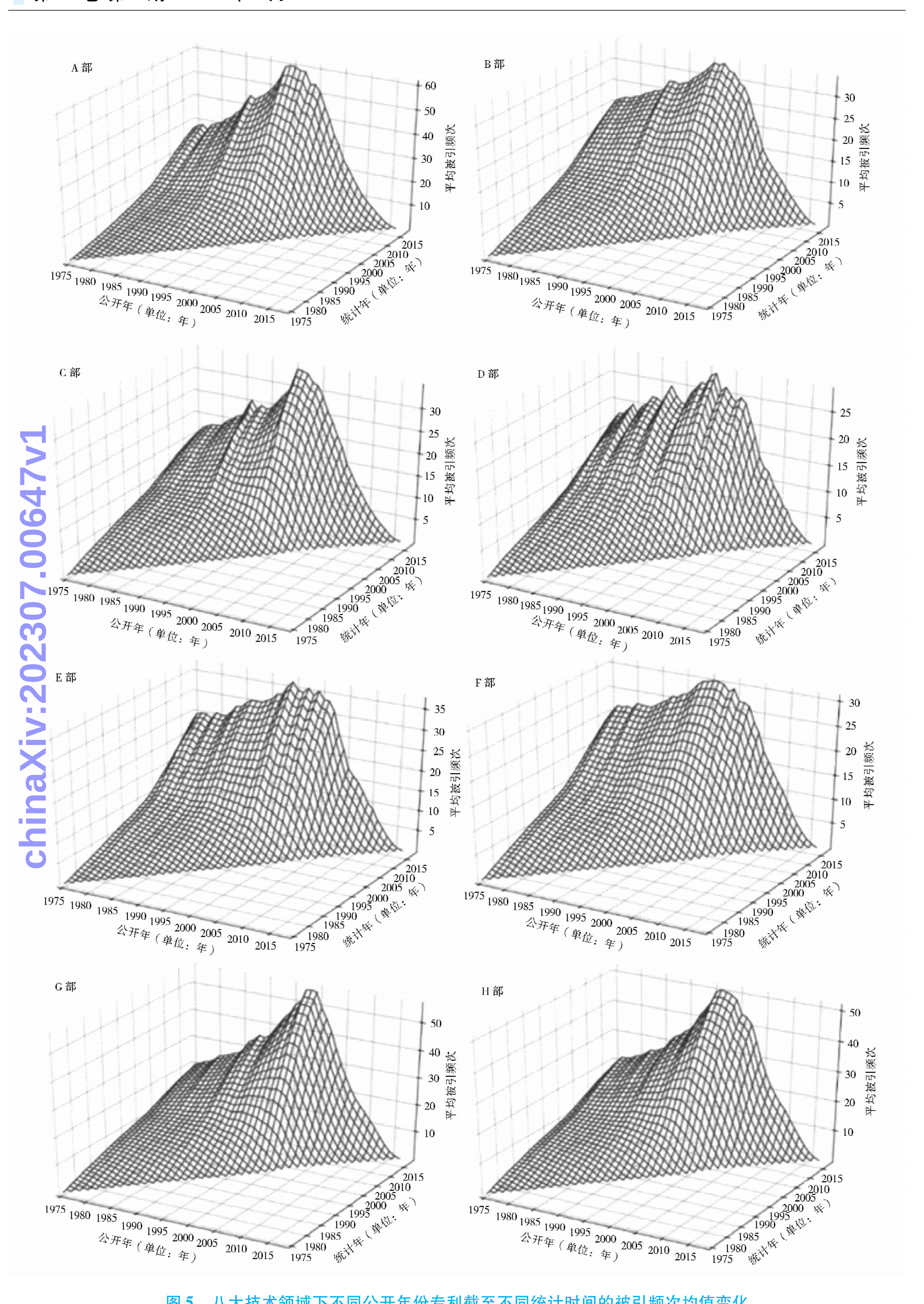



图 5　八大技术领域下不同公开年份专利截至不同统计时间的被引频次均值变化

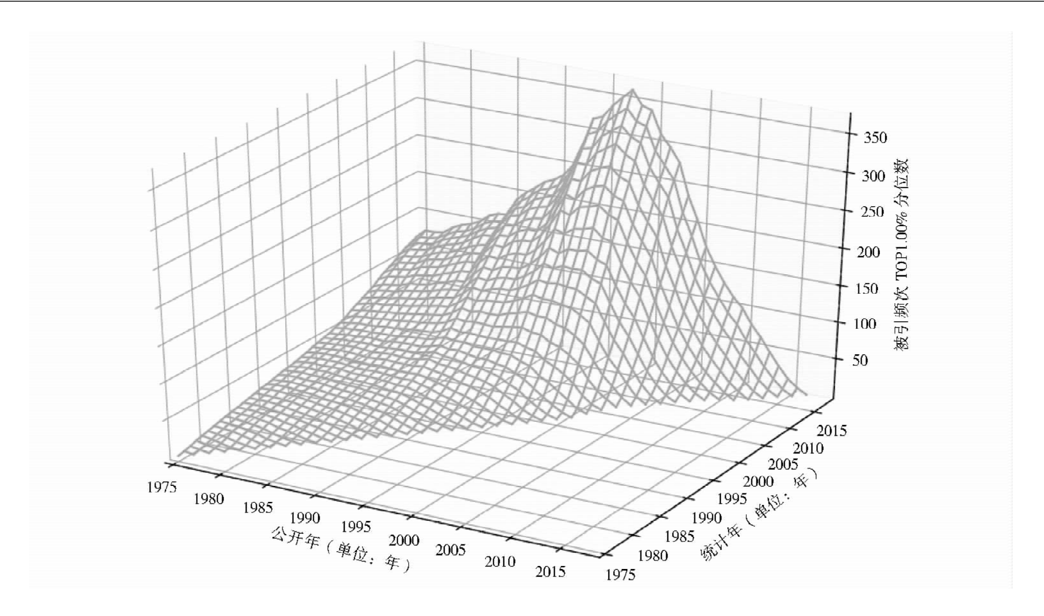



图 6　不同公开年份专利截至不同统计时间的被引频次 TOP1.00%分位数示例



注:图中字母代表 IPC 部,含义分别为:A-人类生活必需;B-作业、运输

图 7　不同技术领域下不同公开年份专利截至不同统计时间的被引频次 TOP 1.00%分位数示例

3.2.2　基于 BP 神经网络模型的基准线预测　本文采用 BP 神经网络模型,预测未来统计时间点的专利被引频次基准线。BP(Back Propagation)神经网络模型由美国的 PDP(Parallel Distributed Processing)研究小组提出。网络结构如图 8 所示,分为输入层、隐藏层和输出层。训练过程分为前向传播过程和反向传播过程。前向传播过程是:将输入样本提供给输入单元,逐层向前传播输入信号,直到产生输出层的结果。反向传播过程是:先对照实际结果计算输出层误差,将误差反向传播到隐藏层神经元,再根据神经元误差,采用梯度下降算法对连接权值和偏置进行优化。

(1)数据准备。将 3.3.1 中基于历史时序的基准线统计结果作为拟合数据集,按照是否划分 IPC 领域、

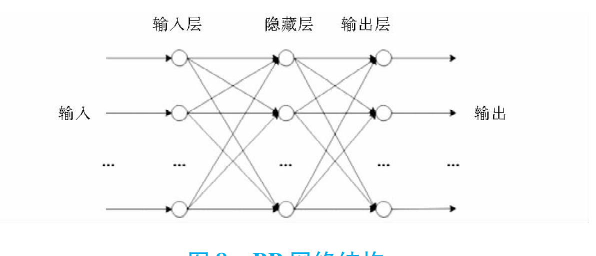



图 8　BP 网络结构

基准值取均值还是 TOP 分位数,将数据集划分成 4 部分:

- 数据集 A。取 1975-2017 年中某年为统计年,计算所有公开年专利截至该统计年时被引频次的均值,共 946 条数据记录。
- 数据集 B。将专利按照 IPC 部划分后,取 1975

-2017 年中某年为统计年,计算所有公开年专利截至该统计年时被引频次的均值,共 7 568 条数据记录。

• 数据集 C。取 1975 -2017 年中某年为统计年,计算所有公开年专利截至该统计年时被引频次的 TOP 分位数,共 946 条数据记录。

• 数据集 D。将专利按照 IPC 部划分后,取 1975 -2017 年中某年为统计年,计算所有公开年专利截至该统计年时被引频次的 TOP 分位数,共 7 568 条数据记录。

(2)模型构建。

• 模型 A。模型 A 的输入数据是数据集 A。输入层设置 4 个神经元,代表公开年、被引公开年、时间间隔和专利数量字段。输出层设置 1 个神经元,代表被引频次均值。隐藏层取 1 层结构。隐藏层神经元数的确定原则为:在满足精度的前提下,取尽可能少的隐藏层神经元数。根据式(1)设置神经元数目。

$$h = \sqrt{m + n} + a \quad \text{式(1)}$$

其中 $h$ 为隐藏层神经元数,$m$ 为输入层神经元数,$n$ 为输出层神经元数,$a$ 为 1 -10 之间的调节常数。模型 A 中,隐藏层神经元个数应该在 4 -12 范围内。在此范围内,设置不同的神经元个数,计算预测误差,找到最佳的隐藏层神经元数。

• 模型 B。模型 B 的输入数据是数据集 B。输入层设置 5 个神经元,代表 IPC 部、公开年、被引公开年、时间间隔和专利数量字段。输出层、隐藏层的设置与模型 A 相同。

• 模型 C。模型 B 的输入数据是数据集 C。预测模型的输入层设置 4 个神经元,代表公开年、被引公开年、时间间隔和专利数量字段。输出层设置 6 个神经元,代表 TOP0. 01%、TOP0. 10%、TOP1. 00%、TOP 10. 00%、TOP20. 00%、TOP50. 00% 分位数。根据公式(1),神经元个数应该为 5 -13 之间。在此范围内,设置不同的神经元个数,计算预测误差,找到最佳的隐藏层神经元数。

• 模型 D。模型 D 的输入数据是数据集 D。输入层设置 5 个神经元,代表 IPC 部、公开年、被引公开年、时间间隔和专利数量字段。输出层、隐藏层的设置与数据集 C 相同。

(3)模型训练。模型训练采用 GroupKFold(分组 k 折验证)方法,将 4 个数据集分别按统计年份分组,同组的样本会同时分给训练集或者测试集。设置 K 为 10,将数据集分成 10 份,轮流将其中 9 份作为训练数据,1 份作为测试数据,进行实验。

(4)模型评价指标。本文选取 MAE(mean absolute error,平均绝对误差)和 MSE(mean squared error,平均平方误差)两个指标对模型进行评估。MAE 按公式(2)求解,MSE 按公式(3)求解。

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|f_i - y_i| \quad \text{式(2)}$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(f_i - y_i)^2 \quad \text{式(3)}$$

(5)模型评估及预测。

• 模型 A。固定隐藏层神经元数量为 4,调整训练次数 epochs,记录不同训练次数的预测误差,结果见图 9。由 MAE 和 MSE 变化曲线可知,训练次数为 100 时,拟合效果趋于稳定。设置训练次数为 100,隐藏层神经元个数分别为 4、5、6、7、8、9、10、11、12,发现当隐藏层神经元个数为 12 时,MAE 和 MSE 值最小(见图 10)。作为参考示例,图 9 和图 10 分别给出不同训练次数的预测误差曲线和隐藏层不同神经元数目的预测误差曲线;后文将不再显示其余模型的曲线变化图。模型 A 中,最终设置训练次数 epochs 为 100,隐藏层神经元个数为 12。

模型 A 的最佳性能曲线如图 11 所示。横轴代表样本的统计年份,纵轴代表被引频次均值。蓝色为实际值,红色为预测值,绿色为两者的差值。从性能曲线来看,整体拟合效果较好。

由模型 A 可以预测截至未来某统计年份时,不同公开年专利的被引频次均值。图 12 是一个预测实例,表示统计年份为 2018 年时,1975 -2017 年公开专利的被引频次均值预测曲线。横轴代表专利公开年份,纵轴代表被引频次均值。

• 模型 B。固定隐藏层神经元数量为 4,调整训练次数 epochs,记录不同训练次数的预测误差。由 MAE 和 MSE 变化情况得到,训练次数达到 500 时,拟合效果趋于稳定。设置训练次数为 500,隐藏层神经元个数分别为 4、5、6、7、8、9、10、11、12,发现当隐藏层神经元个数为 12 时,MAE 和 MSE 值最小。模型 B 中,最终设置训练次数 epochs 为 500,隐藏层神经元个数为 12。

模型 B 的最佳性能曲线如图 13 所示。横轴代表不同技术领域下样本的统计年份,纵轴代表被引频次均值。蓝色为实际值,红色为预测值,绿色为两者的差值。性能曲线显示,被引频次均值的变化幅度小的技术领域(如 B 部、D 部等),拟合效果更好;变化幅度较大的技术领域(如 A 部),拟合效果还需增强。

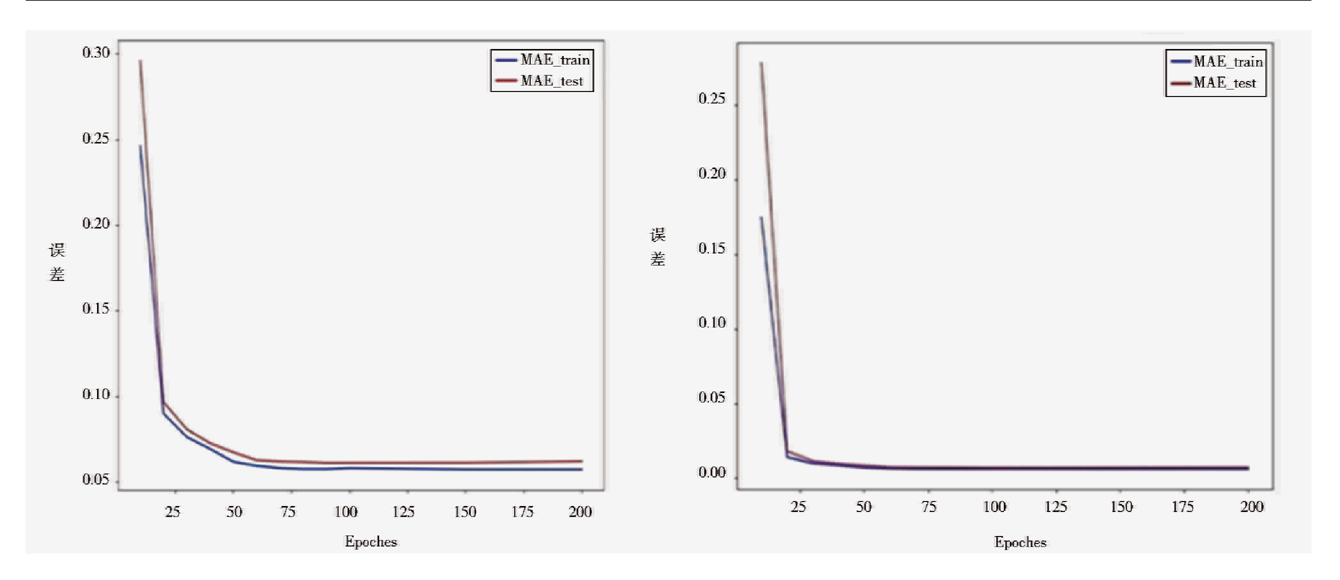



图 9　模型 A 中不同训练次数的预测误差

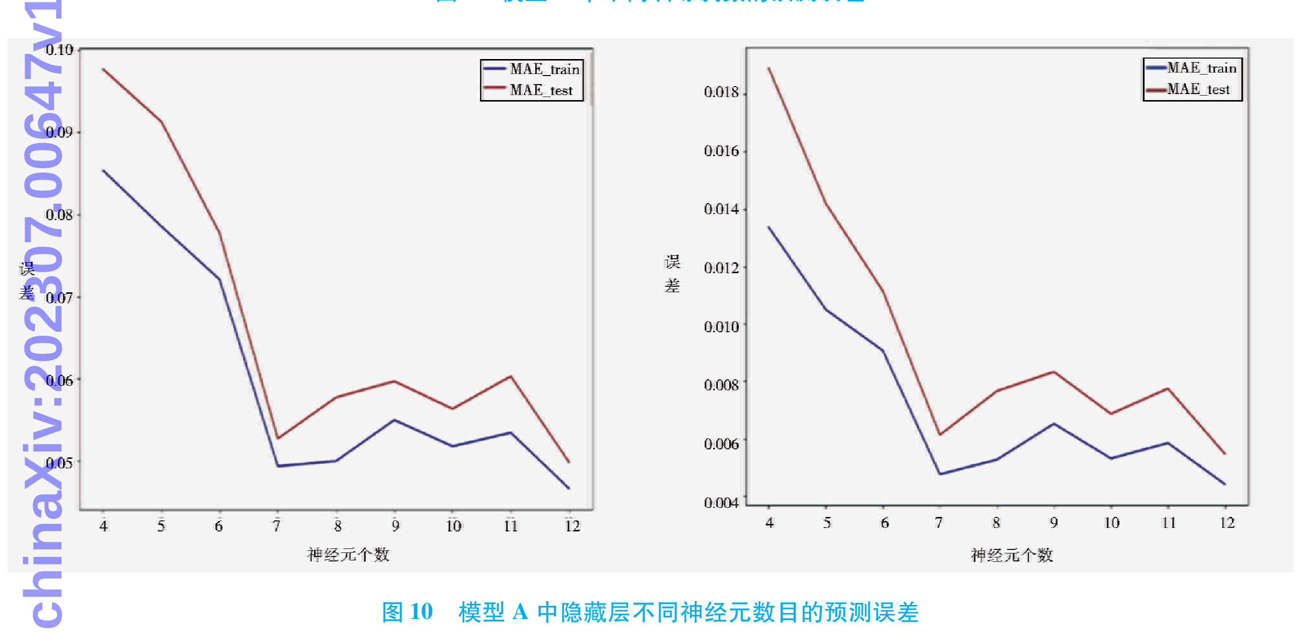





图 10　模型 A 中隐藏层不同神经元数目的预测误差

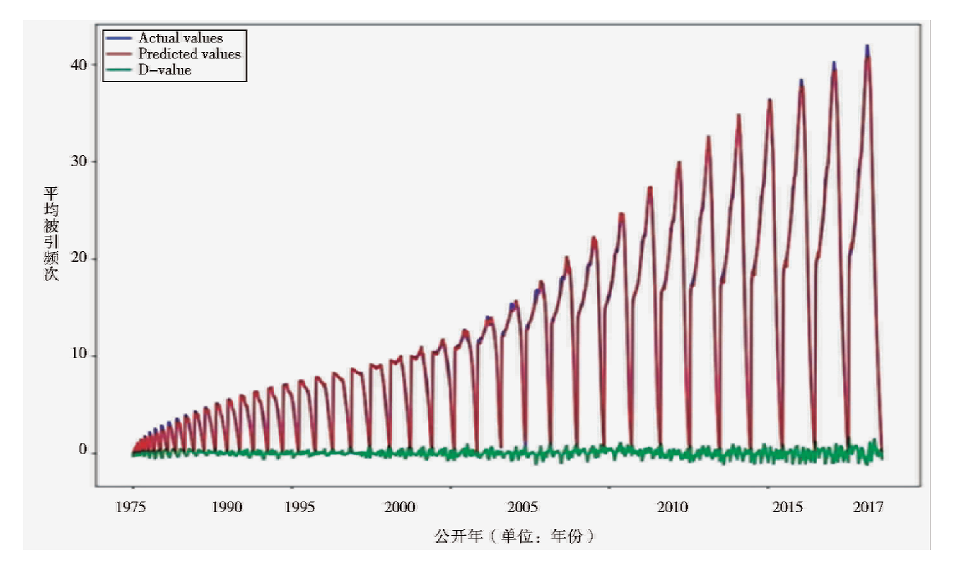



图 11　模型 A 的最佳性能曲线

图书情报工作
第 63 卷 第 9 期　2019 年 5 月

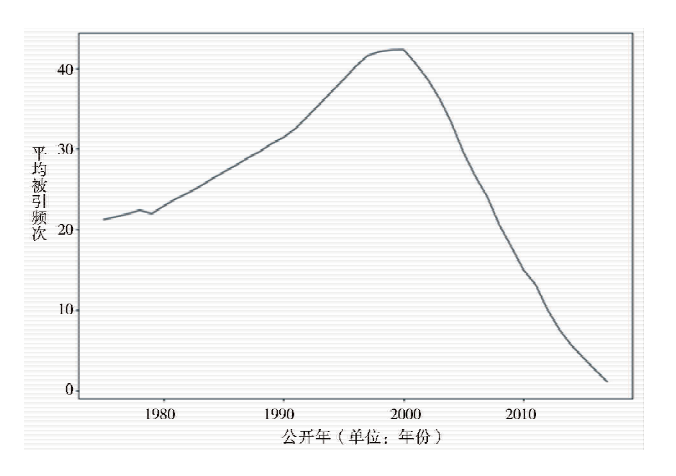



图 12　统计年份为 2018 年时 1975 – 2017 年公开专利的被引频次均值预测曲线

由模型 B 可以预测截至未来某统计年份时，不同技术领域下、不同公开年专利的被引频次均值。图 14 是一个预测实例，表示统计年份为 2018 年时，不同技术领域下、1975 – 2017 年公开专利的被引频次均值预测曲线。横轴代表专利技术领域与公开年份，纵轴代表被引频次均值。

- 模型 C。固定隐藏层神经元数量为 5，调整训练次数 epochs，记录不同训练次数的预测误差。由 MAE 和 MSE 变化情况得到，训练次数设置为 500 及以上时，拟合效果趋于稳定。设置训练次数为 500，隐藏层神经元个数分别为 5、6、7、8、9、10、11、12、13，发现当隐藏层神经元个数为 13 时，MAE 和 MSE 值最小。模型 C 中，最终设置训练次数 epochs 为 500，隐藏层神经元个数为 13。

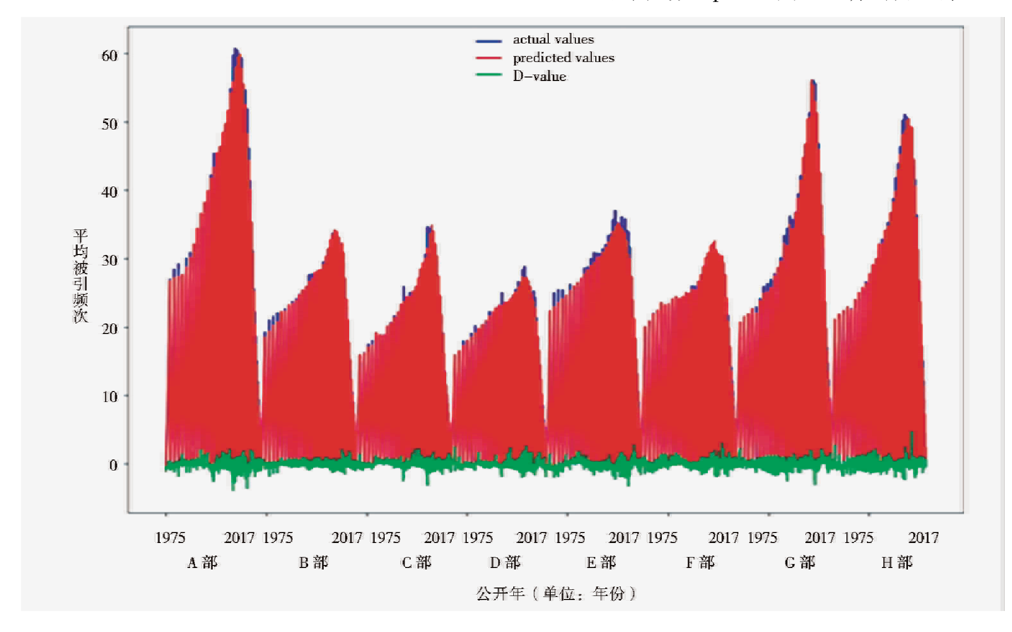



图 13　模型 B 的最佳性能曲线

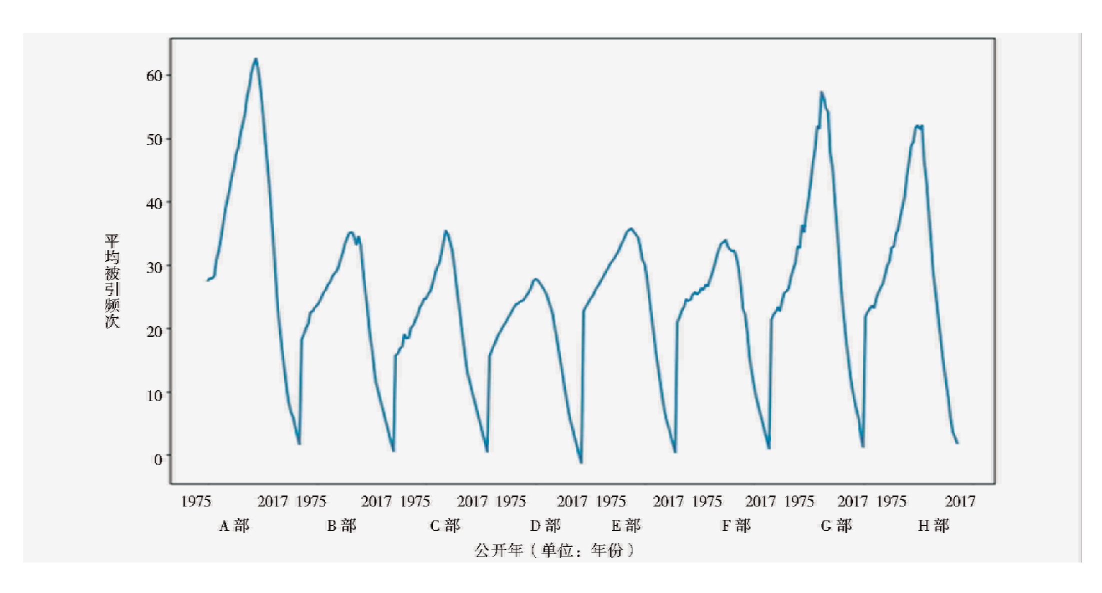



图 14　统计年份为 2018 年时，不同技术领域下 1975 – 2017 年公开专利的被引频次均值预测曲线

模型 C 的最佳性能曲线的示例如图 15 所示。横轴代表样本的统计年份，纵轴代表被引频次的 TOP 1.00% 百分位。蓝色为实际值，红色为预测值，绿色为两者的差值。

由模型 C 可以预测截至未来某统计年份时，不同公开年专利的被引频次 TOP 分位数。图 16 是一个预测实例，表示统计年份为 2018 年时，1975－2017 年公开专利的被引频次 TOP 分位数的预测曲线。横轴代表专利公开年份，纵轴代表被引频次 TOP 分位数。

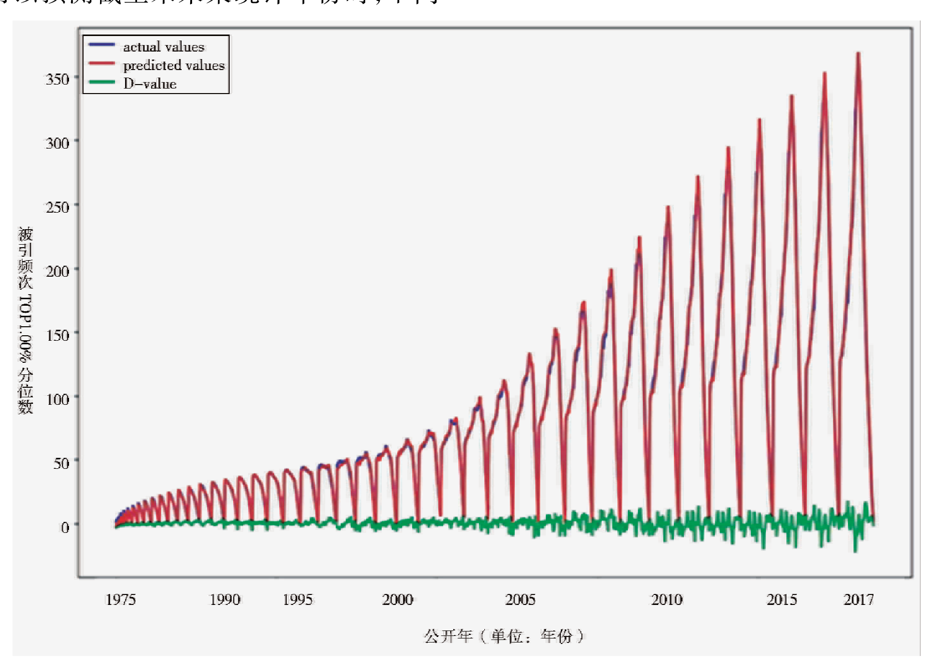



图 15　模型 C 的最佳性能曲线示例（TOP1.00%分位数）

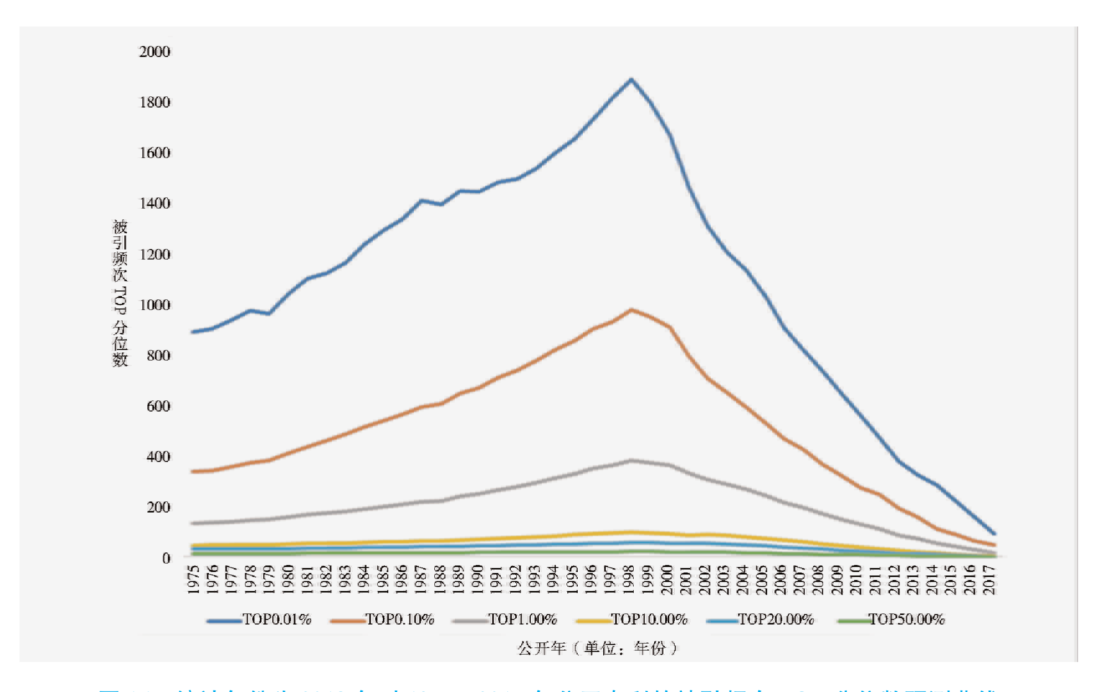



图 16　统计年份为 2018 年时 1975－2017 年公开专利的被引频次 TOP 分位数预测曲线

• 模型 D。固定隐藏层神经元数量为 5，调整训练次数 epochs，记录不同训练次数的预测误差。由 MAE 和 MSE 变化情况得到，训练次数达到 300 时，拟合效果趋于稳定。设置训练次数为 300，隐藏层神经元个数分别为 5、6、7、8、9、10、11、12、13，发现当隐藏层神经元个数为 12 时，MAE 和 MSE 值最小。模型 D 中，最终设置训练次数 epochs 为 300，隐藏层神经元个数为 12。

模型D的最佳性能曲线示例如图17所示。横轴代表不同技术领域下样本的统计年份，纵轴代表被引频次的TOP1.00%百分位。蓝色为实际值，红色为预测值，绿色为两者的差值。性能曲线显示，拟合效果因技术领域不同而有差异，B部、C部等技术领域的拟合效果较好，A部、D部和E部的拟合效果还需增强。

由模型D可以预测截至未来某统计年份时，不同技术领域下、不同公开年专利的被引频次的TOP分位数。图18是一个预测实例，表示统计年份为2018年时，不同技术领域下1975－2017年公开专利的被引频次的TOP分位数的预测曲线。横轴代表专利技术领域与公开年份，纵轴代表被引频次的TOP分位数。

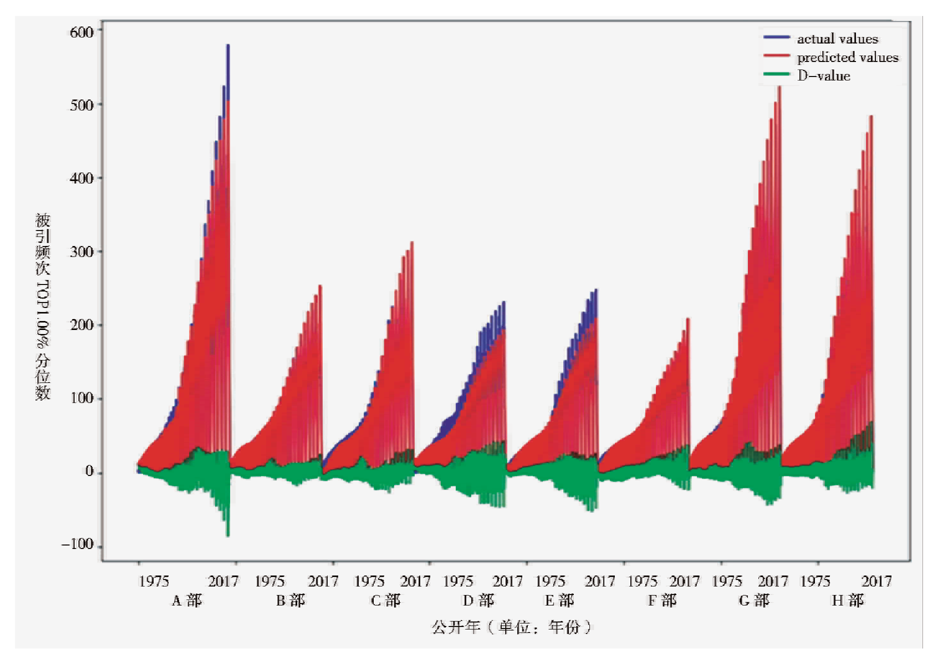



图17　模型D的最佳性能曲线示例(TOP1.00%分位数)

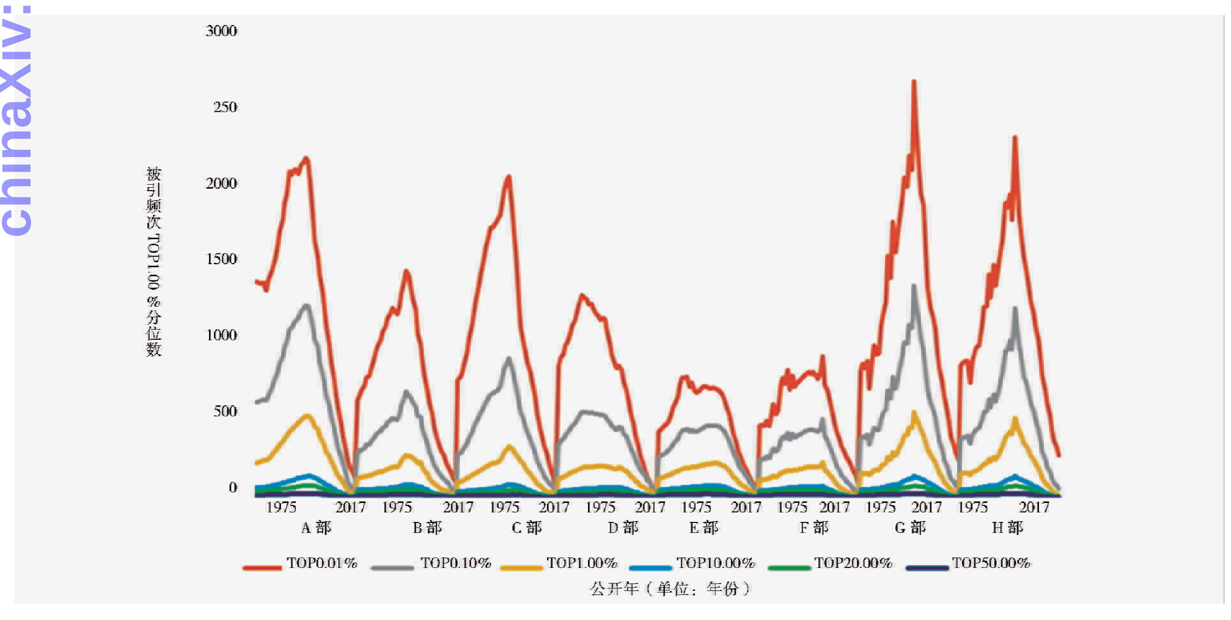



图18　统计年份为2018年时，不同技术领域下1975－2017年公开专利的被引频次TOP分位数预测曲线

## 4　总结与展望

专利引文分析在技术评价活动中应用广泛，专利被引频次是专利引文分析的重要指标。专利被引频次受到时间因素的影响，使得实际评价活动中，被引频次的评价无据可依。基于这一现状，本文开展了时间因素对专利被引频次的影响研究。梳理相关文献发现，时间因素对被引频次的影响，表现为三个方面：①不同年份公开专利的被引频次，无法直接进行比较；②计算

机技术的快速进步,提高了专利审查员的检索能力,导致专利被引机会逐年增大;③只能计算截至统计时间点的专利被引频次,评价年轻专利时误差较大。这些问题的本质在于专利的不同公开年份和不同统计年份的时间差异导致了被引频次无法直接比较。本文从当前统计时间点和历史时序两个角度出发,采用固定效应方法对专利被引频次进行修正。选取被引频次均值和6个TOP百分位数为固定效应水平,建立基于不同技术领域、不同公开年份、不同统计年份的被引频次基准线。构建BP网络模型,对基准线的历史时序变化情况进行拟合,训练得到最优模型。利用最优模型,可以预测未来时间点的被引频次基准线。文中已经提供了截至2018年底统计时,被引频次的预测基准线。研究过程中进一步发现,固定公开年不变时,全领域专利被引频次均值随着统计年的增长呈现上升趋势;固定统计年不变时,全领域专利被引频次均值随公开年的增长呈先增后减的趋势。在不考虑统计时间截断的条件下,专利公开年份越近,全领域专利的被引频次均值随统计时间增加的增长速率也越快。被引频次均值的变化在一定程度上反映了总体的一般水平,恰好验证了前文提出的时间因素对被引频次的影响的三个方面。

本文的创新点:①国内外相关研究多关注论文引文测度指标的时间影响,而本文关注专利引文测度指标的时间影响,并借鉴论文的修正方法进行修正研究,在选题和研究思路上具有新颖性。②在专利领域缺乏不同时间、不同行业的专利文献引文测度基准。本文建立了基于不同领域、不同时间范围的专利被引频次指标基准,使得相关指标的评价有据可依。③前文提到过B. H. Hall的研究开创了专利引文指标时间因素研究的先河。遗憾地是,该研究的数据测度时间截至1999年,而1999年至今又产生了大量的专利引文记录,故该研究无法为当前专利引文指标的合理使用提供参考基准。其他后续研究也没有提出解决方案。针对专利被引频次的基准值逐年变化且难以计算的问题,本文尝试基于时序变化而不是单一时间截面,探究专利被引频次的基准值的逐年变化情况和增长规律,采用神经网络模型拟合并预测未来时间点的基准值。

存在的问题及未来的研究方向:①专利引文分析包含很多有价值的统计指标,而本文仅仅关注专利被引频次指标。未来可以继续探讨时间因素对其他指标的影响。②本文提供了专利被引频次的基准线,但是尚未将成果运用到实际工作中。后续可以开展实证研究,根据实际场景,对时间影响进行修正。通过比较修正前后的结果,探讨本文研究成果的适用性。③文中建立神经网络模型对专利被引频次基准线的历史变化规律拟合,该模型可用于预测未来时间点的被引频次基准线。未来会进一步优化对基准线的拟合预测,尝试建立数学公式或其他直观的方式来表示其变化规律。



**参考文献:**

[1] JAFFE A B, DE RASSENFOSSE G. Patent citation data in social science research: overview and best practices[J]. Journal of the Association for Information Science & Technology, 2016, 68(6): 1360－1374.

[2] 任胜利,王宝庆,郭志明,等.应慎重使用期刊的影响因子评价科研成果[J]. 科学通报,2000,45(2): 218－222.

[3] COZZENS S E. Comparing the sciences: citation context analysis of papers from neuropharmacology and the sociology of science[J]. Social studies of science, 1985, 15(1): 127－153.

[4] HALL B H, JAFFE A B, TRAJTENBERG M. The NBER patent citations data file: lessons, insights and methodological tools[R]. Cambridge: National Bureau of Economic Research, 2001: 25－35.

[5] 万小丽. 专利质量指标中"被引次数"的深度剖析[J]. 情报科学, 2014,32(1):68－73.

[6] BREITZMAN A, THOMAS P. Inventor team size as a predictor of the future citation impact of patents[J]. Scientometrics, 2015, 103(2):1－17.

[7] SCHUBERT A, BRAUN T. Relative indicators and relational charts for comparative assessment of publication output and citation impact [J]. Scientometrics,1986,9(5): 281－291.

[8] GLANZEL W, THIJS B, SCHUBERT A, et al. Subfield-specific normalized relative indicators and a new generation of relational charts: methodological foundations illustrated on the assessment of institutional research performanc [J]. Scientometrics, 2009, 78(1): 165－188.

[9] 陈子凤, 官建成. 国际专利合作和引用对创新绩效的影响研究[J]. 科研管理, 2014, 35(3):35－42.

[10] 蔡虹, 吴凯, 孙顺成. 基于专利引用的国际性技术外溢实证研究[J]. 管理科学, 2010, 23(1):18－26.

[11] WANG J. Citation time window choice for research impact evaluation[J]. Scientometrics, 2013, 94(3):851－872.

[12] ABRAMO G, CICERO T, D'ANGELO C A. A sensitivity analysis of research institutions' productivity rankings to the time of citation observation[J]. Journal of informetrics, 2012, 6(2):192－201.

[13] LEYDESDORFF L, BORNMANN L, MUTZ R, et al. Turning the tables on citation analysis one more time: principles for comparing sets of documents[J]. Journal of the Association for Information Science & Technology, 2011, 62(7):1370－1381.

[14] Centre for Science and Technology Studies, Leiden University, The Netherlands. Leiden ranking [EB/OL]. [2018－10－10]. http:

//www.leidenranking.com/methodology.aspx.

[15] PERNEGER T V. Relation between online "hit counts" and subsequent citations: prospective study of research papers in the BMJ [J]. BMJ, 2004, 329(7465):546-547.

[16] SHEMA H, BAR-ILAN J, THELWALL M. Do blog citations correlate with a higher number of future citations? research blogs as a potential source for alternative metrics[J]. Journal of the Association for Information Science & Technology, 2014, 65(5):1018-1027.

作者贡献说明:
罗文馨:文献调研、研究实施及论文撰写;
赵亚娟:论文选题、修改与审定,研究方案指导。

## Time Impact Study of Patent Cited Frequency

Luo Wenxin[1,2] Zhao Yajuan[1,2]

[1] National Science Library, Chinese Academy of Sciences, Beijing 100190

[2] Department of Library, Information and Archives Management, School of Economics and Management, University of Chinese Academy of Sciences, Beijing 100190

**Abstract:** [**Purpose/significance**] To study the influence of time factor on patent cited frequency can reduce the restriction of time factor on technical evaluation activities and improve the accuracy and reliability of evaluation. [**Method/process**] This paper collected U.S. patent data from 1975 to 2017, and carried out the revision study of patent cited frequency based on fixed effect method. The patents were grouped according to different publication years and different technical fields. Group mean and six TOP quantiles were selected as the benchmarks of patent cited frequency, and the benchmarks of patent cited frequency for both the current time point and the historical time series were counted. Then a neural network model was established to fit the timing variation of the benchmarks, thus predicting the benchmarks of future statistical time points. [**Result/conclusion**] The time difference between publication years and statistical years of patents makes it impossible to directly compare patent citations. This paper establishes benchmarks for patent citations based on different technical fields, different publication years and different statistical years, providing reference for patent evaluation.

**Keywords:** patent cited frequency time fixed effect timing variation benchmark



## 下 期 要 目

□ 在线健康社区老年用户健康信息需求实证研究 (徐孝婷 赵宇翔 朱庆华)

□ 社会化问答平台提问回复率的预测研究——以"百度知道"为例 (邓胜利 付少雄 刘瑾)

□ 5 GAP 模型视角下图书馆营销策略探究 (张若楠 贾苹)

□ 基于指数随机图模型的专利引用关系形成机制研究——以奈拉滨药物为例 (杨冠灿 刘占麟 李纲)

□ 基于网络招聘文本挖掘的课程知识模型自动构建研究 (俞琰 陈磊 赵乃瑄)

□ 基于 kano 模型的高校图书馆微信公众号服务内容分类和供给优先序研究 (李梦楠 周秀会)